PySpark Advanced

讲师：轩宇

DataFrame

PySpark

- Uesr Action Analyser
- PySpark Basics
- SparkSQL
- Case：DSL&SQL
- User-defined functions

SPARK SQL/DATAFRAME

Spark SQL Overview

- Spark module for *structured data* processing (e.g. DB tables, JSON files)
- Three ways to manipulate data:
  - DataFrames API
  - SQL queries
  - Datasets API

DataFrames

- Distributed collection of data organized into *named columns*
- Conceptually equivalent to a table in **relational DB** or **data frame in R/Python**
  - rows, columns, and schema
- API available in Scala, Java, Python, and R

SQL CONTEXT AND HIVE CONTEXT

SQLContext

- Entry point into all functionality in Spark SQL
- All you need is SparkContext

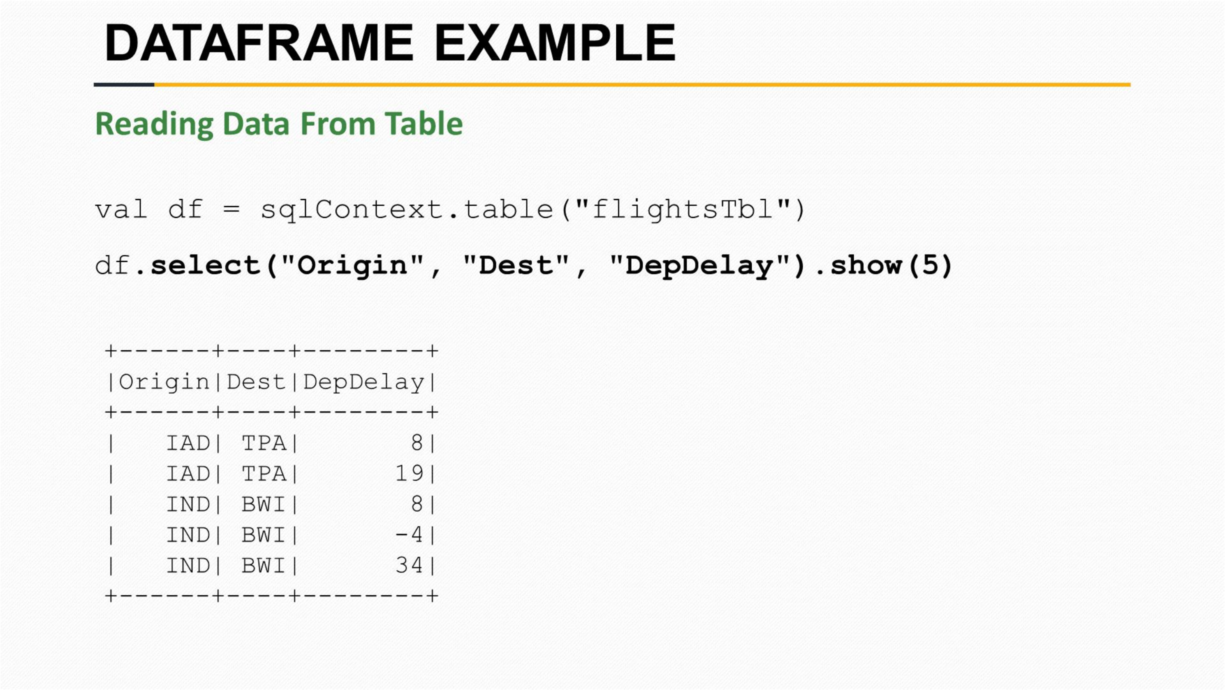
```
val sqlContext = SQLContext(sc)
```

HiveContext

Use when your data resides in Hive

- Superset of functionality provided by basic SQLContext
  - Read data from Hive tables
  - Access to Hive Functions → UDFs

```
val hc = HiveContext(sc)
```

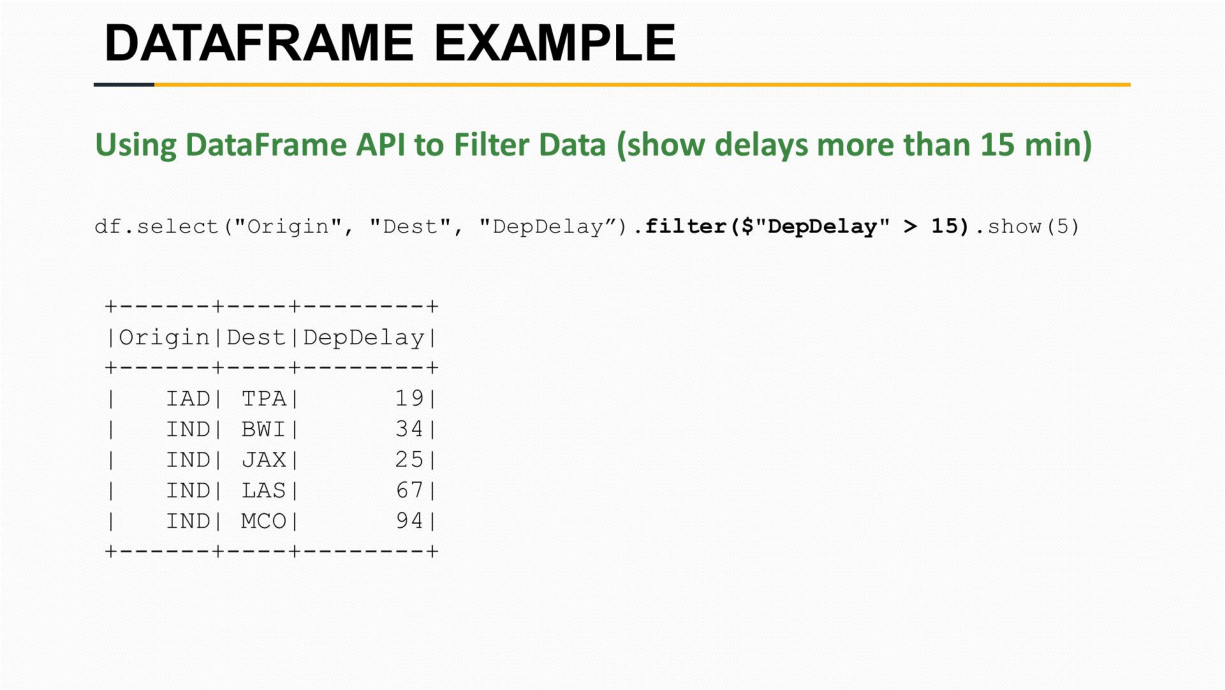


DATAFRAME EXAMPLE

Reading Data From Table

```
val df = sqlContext.table("flightsTbl")

df.select("Origin", "Dest", "DepDelay").show(5)
```

```
+------+----+--------+
|Origin|Dest|DepDelay|
+------+----+--------+
|   IAD| TPA|       8|
|   IAD| TPA|      19|
|   IND| BWI|       8|
|   IND| BWI|      -4|
|   IND| BWI|      34|
+------+----+--------+
```



DATAFRAME EXAMPLE

Using DataFrame API to Filter Data (show delays more than 15 min)

```
df.select("Origin", "Dest", "DepDelay”).filter($"DepDelay" > 15).show(5)
```

```
+------+----+--------+
|Origin|Dest|DepDelay|
+------+----+--------+
|   IAD| TPA|      19|
|   IND| BWI|      34|
|   IND| JAX|      25|
|   IND| LAS|      67|
|   IND| MCO|      94|
+------+----+--------+
```

SQL EXAMPLE

Using SQL to Query and Filter Data (again, show delays more than 15 min)

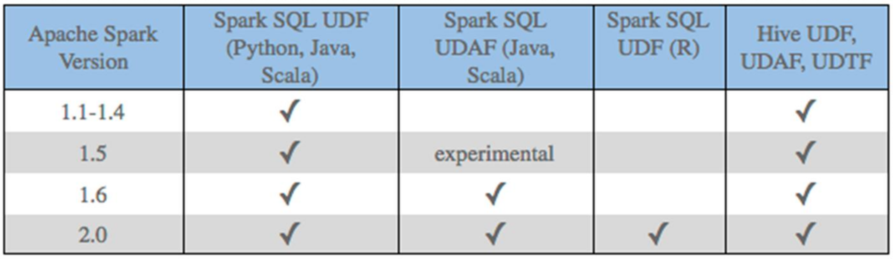
```
// Register Temporary Table

df.registerTempTable("flights")


// Use SQL to Query Dataset

sqlContext.sql("SELECT Origin, Dest, DepDelay
                FROM flights
                WHERE DepDelay > 15 LIMIT 5").show
```

```
+------+----+--------+
|Origin|Dest|DepDelay|
+------+----+--------+
|   IAD| TPA|      19|
|   IND| BWI|      34|
|   IND| JAX|      25|
|   IND| LAS|      67|
|   IND| MCO|      94|
+------+----+--------+
```

USER-DEFINED FUNCTIONS

- Functions available for DataFrame
  - **org.apache.spark.sql.functions**
- Functions for registering user-defined functions
  - **SQLContext.udf.register(name, func)**

```
sqlContext.udf.register("myUDF", (arg1: Int, arg2: String) => arg2 + arg1)
```



Apache Spark Version	Spark SQL UDF (Python, Java, Scala)	Spark SQL UDAF (Java, Scala)	Spark SQL UDF (R)	Hive UDF, UDAF, UDTF
1.1-1.4	✓			✓
1.5	✓	experimental		✓
1.6	✓	✓		✓
2.0	✓	✓	✓	✓